Zusammenfassung der Dissertation

Visual Data Mining of Graph-Based Data

Oliver Niggemann

Thema der Arbeit ist die Analyse und Visualisierung von als Graph gegebener Daten. Durch die Anwendung von maschinellen Lernverfahren, Wissensakquisitionsmethoden und Graph Layout Algorithmen wird der Benutzer in die Lage versetzt, auch große und komplexe Graphen zu verstehen und die in ihnen enthaltene Information zu extrahieren.

Zu Beginn der Arbeit werden wünschenswerte Eigenschaften von Verfahren zur Analyse und Visualisierung von Graphen definiert. Mittels dieser Eigenschaften erfolgt eine neue Klassifikation von existierenden Methoden zur Informationsvisualisierung. Dieselbe Klassifikation dient auch zur Bewertung der im Rahmen der Dissertation neu entwickelten Verfahren.

Ausgehend von oben erwähnten wünschenswerten Eigenschaften, wird eine neue allgemeine Methodik zur Graphanalyse und Graphvisualisierung präsentiert. Die einzelnen Schritte dieser Methodik werden im ersten Teil der Arbeit im Detail erläutert; im zweiten Teil wird die Methodik benutzt, um Lösungen für Analyse- und Visualisierunsprobleme aus sechs verschiedenen Anwendungsgebieten zu finden. Einzelschritte und Anwendungen werden jeweils diskutiert und bewertet.

Der erste Schritt der Methodik benutzt Clustering-Verfahren zur Strukturanalyse von Graphen. Die Arbeit klassifiziert bestehende Verfahren nach Menge und Art des nötigen Wissens und nach dem verwendeten Optimierungskriterium.
Der Autor stellt drei neue Clustering-Methoden vor:

1. MajorClust ist ein schnell zu implementierendes und performantes Verfahren, welches bereits erfolgreich in mehreren Domänen angewandt wurde. Es werden theoretische und empirische Laufzeitabschätzungen sowie Analysen der Clustering-Qualität gegeben.
2. Das Prinzip der Λ-Optimierung beruht auf der Definition einer Funktion, welche die Güte eines gegebenen Clusterings bewertet. In der Arbeit wird eine neue solche Funktion definiert, das sogenannte Λ-Maß, welches sich deutlich von bisherigen Maßen unterscheidet.
3. Während die ersten beiden Methoden zum Clustering nur den Graphen (d.h. Knoten, Kanten, Kantengewichte) benötigen, ist die Methode „Clustering by Example“ in der Lage, zusätzliches Wissen über das Problem (in der Arbeit am Beispiel von hydraulischen Schaltkreisen dargestellt) in Form von Knoten-Labeln zu verarbeiten. Idee ist es, Beispiel-Clusterings zu erstellen und mittels Methoden des maschinellen Lernens ein Clustering-Verfahren zu abstrahieren.

Die Klassifizierung von Teilgraphen (d.h. Clustern) ist Gegenstand des zweiten Teilschrittes der Methodik. Diese Klassifikation erlaubt es, Cluster durch einen Text zu bezeichnen und die Visualisierung dadurch noch verständlicher zu machen. Die Arbeit diskutiert insbesondere Methoden zur benutzerfreundlichen Definition von Klassifikationswissen. So werden Beispielklassifikationen benutzt, um eine Klassifikationsfunktion zu erlernen. Neben der Parametrisierung von Klassifikationsfunktionen (z.B. durch Regression oder neuronale Netze) werden auch fallbasierte Ansätze besprochen.
Eine weitere in der Arbeit vorgestellte Anwendungsmöglichkeit für Cluster-Klassifikation ist die automatische Auswahl von individuellen Layout-Methoden für jedes Cluster. Dies ermöglicht, unter Berücksichtigungen der Vorlieben des Benutzers, für jeden Teilgraph die optimale Graphlayout Methode einzusetzen.

Im letzten Schritt der Methodik wird der Graph zweidimensional visualisiert, d.h. es werden Positionen für seine Knoten ermittelt. Existierende Verfahren werden mittels eines Klassifikationsschemas eingeordnet.
Die Arbeit erläutert zunächst die Funktionsweise der bekanntesten Methoden und geht auf ihre Vor- und Nachteile ein.
Anschließend wird ein neues Verfahren vorgestellt, welche mittels statistischer Methoden ein Initiallayout findet, wobei insbesondere die globale Struktur des Graphen berücksichtigt wird. Dieses Initiallayout wird lokal durch die Benutzung eines sogenannten „Spring-Embedder“-Algorithmus verbessert.
Die zuvor besprochenen Graph Layout Methoden werden im Rahmen der Arbeit so erweitert, daß sie die ermittelten Strukturinformationen über den Graphen (siehe Clustering) berücksichtigen können.

Am Ende des ersten Teils der Arbeit wird die Umwandlung von als Tabelle gegebener Daten in einen Graph diskutiert. Das hierzu notwendige Expertenwissen über Objektähnlichkeiten wird unter Verwendung von Algorithmen aus dem Bereich des maschinellen Lernens aus abstraktem Benutzerwissen gewonnen, d.h. es wird versucht, das Wissen auf eine für den Experten natürliche Art und Weise zu ermittelt. Neben der Verwendung von existierenden Wissensquellen (z.B. Objektklassifikationen, Einordnungen von Objekten in Taxonomien wie

z.B. hierarchischen Dateisystemen) wird auch eine visuelle Ähnlichkeitsdefinition vorgestellt: Beispielobjekte werden vom Experten visualisiert, aus der Visualisierung wird ein Maß für die Objektähnlichkeit gewonnen.

Im zweiten Teil der Arbeit wird die zuvor vorgestellte Methodik zur Lösung von Problemen aus sechs verschiedenen Anwendungsfeldern benutzt. Hierzu wird das im Rahmen der Arbeit entwickelte Softwaresystem StructureMiner verwandt, welches die im ersten Teil der Arbeit vorgestellten Verfahren umsetzt. Die Funktionalitäten von StructureMiner werden erläutert; neben der Implementierung der drei Schritte Clustering, Klassifikation und Graph Layout umfaßt das System auch die Fähigkeiten eines Graph Editors. StructureMiner bietet dem Benutzer die Auswahl zwischen 5 verschiedenen Clusteringverfahren und 7 Graph Layout Algorithmen.

Die erste besprochenen Anwendung ist die Visualisierung von Verkehr in Rechnernetzen. Hierzu werden Rechner (d.h. IP-Adressen) als Knoten eines Graphen verstanden, während die Kanten des Graphen durch den Verkehr zwischen den Knoten definiert werden. Neben der Anwendung der im ersten Teil besprochenen Methodik auf diese Graphen, stellt die Arbeit auch ein System zur Visualisierung von zeitlichen Verkehrsveränderungen und ein Visualisierungsverfahren zur Unterstützung der Planung von Rechnernetzen vor. Des weiteren werden in der Arbeit die Zusammenhänge zwischen Visualisierung und Simulation am Beispiel eines neu entwickelten Systems zur qualitativen Simulation von Verkehr in Rechnernetzen herausgearbeitet.

Als zweite Anwendung wird die Analyse, Visualisierung und Administration von Wissensbasen zur Konfiguration von technischen Systemen diskutiert. Neben diverser Anpassungen von StructureMiner liegen die Schwerpunkte hier in der (schon oben erwähnten) automatischen Auswahl von individuellen Layoutverfahren für die Cluster und in der automatischen Klassifikation (d.h. Bezeichnung) von Clustern, welche hier technischen Subsystemen entsprechen.

Die dritte Anwendung ist die Visualisierung von Protein-Interaktionsgraphen aus dem „Human Genome Project“. Protein-Interaktionen stellen das Bindeglied zwischen dem nun größtenteils bekannten menschlichen Genom und den biologischen Prozessen im menschlichen Organismus dar. Die Arbeit zeigt, daß die Anwendung der in dieser Arbeit entwickelten Methodik dem Experten bei der Analyse und beim Verständnis dieser Graphen hilft.
Des weiteren wird ein Maß für den Vergleich von Clustering-Ergebnissen und existierenden Protein-Taxonomien vorgestellt. Dies ist u.a. wichtig für die Beurteilung solcher Taxonomien.

Die Visualisierung einzelner fluidischen Schaltkreise ist die vierte Anwendung. Der Schwerpunkt liegt hierbei auf dem Clustering solcher Schaltkreise. Insbesondere die oben erwähnte neue Clusteringmethode „Clustering by Example“ wird benutzt.

Das Dokumentenmanagement technischer Dokumente stellt die fünfte untersuchte Problemstellung und zugleich die erste von zwei Anwendungen dar, bei der die Daten zuerst aus einer tabellarischen Form in einen Graph umgewandelt werden müssen. Hierzu werden die oben erwähnten neuen Wissensakquisitionsmethoden verwandt.
Es wird zuerst die Umwandlung einer Sammlung technischer Dokumente (am Beispiel von fluidischen Schaltungen) in eine Tabelle besprochen. Diese Umwandlung geschieht durch die Anwendung komplexer Verfahren aus dem Bereich der fluidischen Systeme (z.B. Bestimmung fluidischer Achsen, Fahrprofilbestimmung für die Zylinder etc.). In einem zweiten Schritt wird diese Tabelle unter Benutzung der bereits erwähnten Wissensakquisitionsmethoden in einen Graph transferiert, welcher in einem dritten Schritt mittels der in der Arbeit vorgestellten Analyse- und Visualisierungsmethodik dargestellt wird. Das so gewonnen Layout hilft dem Benutzer beim Zugriff auf große Dokumentensammlungen.

Die sechste und letzte Anwendung ist die Abhängigkeitsanalyse von tabellarischen Daten. Hier wird die Abhängigkeit zwischen den Spalten einer Datentabelle (d.h. zwischen Objekteigenschaften) untersucht. Der Autor überführt dazu eine Datentabelle in einen Graph, indem er Spalten zu Knoten macht und die jeweiligen Abhängigkeiten durch Kantengewichte ausdrückt. Zur Bestimmung der Spaltenabhängigkeiten werden Verfahren aus der Statistik (Regression, χ^2-Test etc.) eingesetzt. Anhand einer Datensammlung zur Bevölkerung der U.S.A. wird ersichtlich, daß die Anwendung der vorgestellten Methodik auf solche Abhängigkeitsgraphen die Analyse von komplexen Datensätzen unterstützt.

Zum Abschluß werden die Ergebnisse zusammengefaßt und bewertet. Im Anhang findet sich eine Zusammenfassung der in der Arbeit verwandten Verfahren zum maschinellen Lernen, wobei Gemeinsamkeiten und Unterschiede herausgearbeitet sind.